COMPARISON OF NON-ENSEMBLED AND ENSEMBLED SUPERVISED MACHINE LEARNING TECHNIQUES FOR PM2.5 PREDICTION IN GUWAHATI CITY

Shrabani Medhi, Ph.D Research Scholar, Department of Computer Science and Engineering, Girijananda Chowdhury Institute of Management and Technology, Guwahati, Affiliated to Assam Science and Technology University, India
Dr. Minakshi Gogoi, Associate Professor,Department of Computer Science and Engineering, Girijananda Chowdhury Institute of Management and Technology, Guwahati, Affiliated to Assam Science and Technology University, India

Abstract:

Air pollution is a major concern for Guwahati city. Guwahati, one of the important cities of North-East India, is one of the cities which has one of the highest Black Carbon levels of pollution in the world. Till now only a few researchers have undertaken this major issue of Guwahati. After doing literature review, a strong need is felt to fill this gap by performing systematic analysis and prediction of air pollution in Guwahati city. Out of all the pollutants, PM2.5 is the most dangerous one. A comprehensive analysis is done on the prediction of PM2.5 concentration using non-ensembled supervised machine learning techniques and ensembled learning techniques, namely, bagging, boosting and voting. Preprocessing of the dataset is done and correlation analysis is done to select the key features. Ten cross fold validation method is used along with GridSearchCV and RandomSearchCV for hyperparameter tuning. We have performed a comparative study to determine the best model that accurately predicts PM2.5 concentration. Root Mean Square, Mean Absolute Error (MAE), Mean Squared Error, Root Mean Squared Log Error and R2 score are used as evaluation criteria of the regression models. Models are fitted using hyperparameter tuning to determine the best regression model with respect to error rate. It is concluded that ensemble learning performs better than their corresponding non-ensemble machine learning techniques. Furthermore, weak non-ensemble machine learning technique were transformed into strong ensemble learners when combined with other machine learning techniques to give better prediction performance when appropriate ensemble technique is used.

Key words: Air pollution, PM2.5, ensembled learning, air pollution prediction, bagging, boosting, voting.

1. INTRODUCTION

In recent years due to increased growth, urbanization and improved lifestyle in Guwahati city air pollution have increased tremendously. Guwahati has one of the highest Black Carbon pollution levels in the world (Barman & Gokhale, 2019). The concentration of PM2.5 in Guwahati is much higher than the permissible limit. A high concentration of PM2.5 is extremely dangerous. Paper (Evans et al., 2013; Laden et al., 2006; Pope III & Dockery, 2006) has proven the association of PM2.5 with cardiovascular disease, cancer, respiratory disease, metabolic disease, and obesity. It is very important that the prediction of PM2.5 concentration should be done so that effective measures can be taken beforehand. Many statistical linear methods were used in past to predict air pollution but these methods have major drawbacks like complexity and variation in time-series data (Hsieh et al., 2015; Johnson et al., 2010). We have tried to determine whether non-ensembled or ensembled machine learning algorithm predicts the PM2.5 concentration better.Due to rapid urbanization, industrialization, and high vehicular emissions in Guwahati, air pollution has come up as a major problem. For the past 10-12 years, Guwahati has emerged as one of the rapidly growing cities in India. The Pollution Control Board of Assam (PCBA) has its headquarters in Bamunimaidam, Guwahati which has a central laboratory. It monitors the city’s ambient air quality and has revealed that the city has a PM2.5 presence well above the prescribed value since 2008(Kioumourtzoglou et al., 2016). Guwahati is one of the cities which has the highest concentration of black carbon in the world. Since Guwahati has one of the highest concentrations of black carbon in the world, it is very important that some serious steps are taken in this regard to deal with air pollution. It is very important that the prediction of PM2.5 concentration should be done so that effective measures can be taken beforehand.

PM2.5 are solid and liquid suspended particles in the air. For example, dust, soot, ash, etc. The size of PM2.5 has a width of less than two and one-half microns or less. Due to the miniature size of PM2.5, they have the capacity to travel deep into the respiratory tract and reach the lungs. Exposure to particulate matter can result in many short-term health issues. For example, runny nose, eye irritation, nose irritation, throat and lung irritation, cough, sneezing and shortness of breath. Prolonged exposure to PM2.5 may result in serious health issues which affect lung function and deteriorate medical conditions such as asthma and heart diseases(Gonzalez-Gorman et al., 2019; Janssen et al., 2013). Adults with chronic lung and heart diseases, children, infants, and asthmatics are more likely to experience adverse health issues due to high PM2.5 concentrations exposure (Kim et al., 2015). Children inhale more air per kg of body weight than adults. They breathe faster, have smaller body sizes, and spend more time outdoors than indoors. So, children are more prone to diseases due to PM2.5 concentration. According to World Health Organization (WHO), there is an approximate estimation that globally 3.0% of cardiopulmonary and 5.0% of lung cancer deaths are caused due to particulate matter exposure(Kim et al., 2015). PM2.5 is highly more toxic compared to other air pollutants because it contains nitrates, sulfates, metals, acids, and particles with various chemical compositions that are adsorbed on their surfaces. They can be penetrated easily into the indoor areas and can be transported over a much longer distance. All these issues make PM2.5 very dangerous. PM2.5 is monitored by the monitoring stations and Air Quality Index (AQI) is calculated based on it. Air Quality Index is a number used by the government to show to the public the quality of the air. An increase in AQI represents increase in air pollution. According to Indian government (CPCB), there are six categories of air quality namely, good, satisfactory, moderate, poor, very poor and severe. Eight major air pollutants are taken for AQI calculations namely, particulate matter (PM10, PM2.5), ozone (O_3), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ammonia (NH3) and lead (PB). It is shown in Table 1(Pope III & Dockery, 2006).

The formula to calculate AQI as per Indian CPCB is given as (Sharma, 2021)

$$\mathrm{AQI} = \frac{[\mathrm{Ihigh} - Ilo\ \ ]}{[BPhigh - BPlow]} * (\mathrm{CP} - \mathrm{BPlow}) + \mathrm{Ilow} \qquad (1)$$

where,
AQI = Air Quality Index, CP = Pollutant Concentration
BPhigh = Concentration breakpoint that is >=CP, BPlow= Concentration breakpoint that is <CP,
Ihigh= AQI value corresponding to BPhigh, Ilow = AQI value corresponding to BPlow

Air quality indices is calculated for each of the pollutant separately and the highest of all the pollutants values gives a location's AQI at a given point of time. Various models have been reviewed in the literature to predict air pollution starting from statistical, deterministic, physical, machine learning and various deep learning techniques. The traditional models are based on statistics and probability which are very complex and yet less efficient. Air pollution prediction using ML have proved to be more reliable, consistent and efficient. Data collection has been made very easy due to the advanced technologies and sensors. ML algorithms can perform accurate and reliable predictions through extremely large pollution data which requires rigorous analysis. In this paper we have investigated 2 years of hourly air pollution data of Guwahati city and have analyzed meteorological conditions, criteria gases and particulates. Preprocessing and cleaning of data is done first. Better insights are developed and hidden patterns and trends are investigated using data visualization methods. In paper (Medhi &Gogoi, 2021), various visualization techniques are discussed to analyze the air pollution data. Correlation coefficient with ML is used which is used by very few researchers in literature(Alade et al., 2019). Resampling technique is used to identify and address data imbalance. Using the resampling technique, the non-ensembled and ensembled ML techniques are exercised. Their performance is analyzed and compared using standard metrics. In this paper, we have done prediction of PM2.5 using ensembled and non-ensembled supervised machine learning techniques that are described in Section 3. We have done a comparative analysis of these models using RMSE, MAE and R2 as the evaluation criteria. We have also analyzed these models by fitting hyperparameter tuning in Jupyter. In this paper we have tried to propose the optimal model based on error rate. A review of the literature review is given in Section 2. System Evaluation is given in Section 3. Methodology is described in Section 4. Results and Discussions are addressed in Section 5. Conclusion is presented in Section 6.

2. LITERATURE REVIEW

In past few years many machine learning methodologies has been proposed to solve problems related to air pollution. In this section we have done a thematic literature review to present some of the key work done in this field. Further, we have done a review on why PM2.5 is considered to be so dangerous. A machine learning technique was used to perform the daily air pollution prediction of 74 cities in China (Xi et al., 2015). WRF-Chem models and five different classification models were used. The results showed that ANN has the drawback of a low convergence rate. In (J. Zhang & Ding, 2017), Extreme Learning Machine (ELM) was applied to perform prediction on Hong Kong data. The algorithm performs good in relation to precision, generalization and robustness. RMSE of 95 and training time of 0.07s was achieved. In (Asgari et al., 2017), Logistic Regression and Naïve Bayes algorithm was used to predict air pollution. They analyzed data from 2009 to 2013 in Tehran using Apache Spark. Naïve Bayes showed good results but it is not appropriate for predicting real time series data.

Table 1: AQI Values of PM2.5

Category of AQI	**Air Quality Index**	
	Index Value	***Breakpoints for PM2.5 (μ/m^3, average for 24 hours)***
Good	Between 0 to 50	Between 0.0 to 30.0
Satisfactory	Between 51 to100	Between 31 to 60
Moderate	Between 101 to 200	Between 61 to 90
Poor	Between 201 to 300	Between 91 to 120
Very poor	Between 301 to 400	Between 121 to 250
Severe	Between 401 to 500	250+

A hybrid machine learning system (HISY-COL) is suggested in paper (Bougoudis et al., 2016). This method tries to identify the correlation between air pollutant levels and weather patterns. One of the goals is to find the cause of the air pollutants. Artificial Neural Network (ANN) and Random Forest is used which increases the accuracy. Drawback of the system is that feed-forward neural network fails to predict continuous values. Training data is also limited. In paper (Yan et al., 2017), an attempt is made to increase the accuracy by applying neural network in two phases. Meteorological parameters are trained and then used for air pollutant analysis. The main drawback is that only one station is considered with few hours of data. Neural network has the problem of overfitting when small datasets are used. In (Li et al., 2016), a comparison is made between neural network with auto regression moving average (ARMA) and support vector regression (SVR) models. The accuracy for neural network is found to be better. However, processing time of the models is not mentioned. Wind speed is responsible for ventilating air pollutants and transporting them to other areas even though emission sources may not be present in those regions (Wang & Ogawa, 2015; F. Zhang et al., 2014).

Increased relative humidity tend to make particulate matter heavier and helps in dry deposition process of removal. Scavenging by wet deposition is directly related to precipitation.Table 1 presents a comparative and concise analysis of the literary work done for the prediction of air pollution. It has been observed that less attention from scholars is obtained for the prediction and analysis of air pollution for Guwahati city in spite of being one of the most polluted cities of the world. In this paper we have tried to fill this gap by analyzing 2 years hourly air pollution data from Guwahati city. It is an earnest attempt from our side to contribute to the literature with data visualization, exercising correlation coefficient-based statistical outliers for analysis and comparison of the non-ensembled and ensembled ML techniques by using the standard performance metrics.

3. SYSTEM EVALUATION

In this section we have represented what type of system we have used to perform our processing. We have also shown the evaluation criteria that is used to evaluate the methods that we have used. We have conducted all our experiments on i7 machine running on Windows 10 operating system. The preprocessing of data, time series evaluation and all the algorithms are implemented using Python programming and its libraries. We have used matplotlib to plot the graphs. Performance evaluation is done using sklearn metrics. Ten cross fold validation method, GridSearchCV and RandomSearchCV function is used to perform hyperparameter tuning. MAE, RMSE and R2 score is used to evaluate the performance.A model's accuracy is considered to be higher if MAE, RMSE, MSE, RMSLE values are lower. On the other hand, a higher R^2 score is preferred. R^2 score helps to indicate how well the predictor variable can perform with the variation in the response variables. MAE, RMSE, RMSLE and MSE indicates how well a model can predict the value of response variable in absolute terms.

4. METHODOLOGY

We have used non-ensembled and ensembled supervised machine learning techniques to predict the concentration of PM2.5 in the city of Guwahati and have made a comparison to find the best model. A five-step procedure is followed which is shown in Fig 1. The detailed process is explained in the below section. We have implemented non-ensemble machine learning techniques (MLR, DT, SVM, ANN) and ensemble learning techniques (bagging, boosting, voting) to predict PM2.5 concentration and evaluate their accuracy and error metrics.

Data source: The Continuous Ambient Air Quality Monitoring Station (CAAQMS) data for Guwahati city used in this study is collected from Pollution Control Board, Assam located in Bamunimaidam. The data set contains CAAQMS data of Guwahati city from January, 2019 to December, 2020 (2 years). The parameters used in the study are given in Table 3. Meteorological conditions, criteria gases and particulates are used as parameters. Total 9 features are used. Table 4 provides a descriptive statistic of the available meteorological conditions, criteria gases and particulates measures: count, mean, standard deviation, min, 25%, 50%, 75%, maximum, skewness, kurtosis and variance. There is no high value of skewness in data. It indicates that there is no sharp increase in the data. The high value of kurtosis in PM2.5 indicates the presence of data discontinuities. While modeling the data it is important the summary statistics are consistent, in other words, the time series is stationary. We have used three methods to check stationarity.

Time plots: For time series analysis, time plots are very important as they are used as a descriptive tool that may show both seasonality and trend, outliers and discontinuities. This allows us to take better decisions in choosing the appropriate technique to perform the prediction. The time plots of PM2.5 used is shown in Fig 2. The time plots of all the features used is created to check for stationarity. From the plots it can be observed that the distribution for each of the data is non-linear. A time series is stationary if the variance remains same over time. The plot in Figure 2 indicates the stationarity of the data.

Table 2 Literature review on air pollution prediction using ML techniques

Sl No	**Author(s) and Year**	**Algorithm applied**	**Parameters used**	**Other techniques applied**	**Evaluation metrics used**	**Result**
1	(Betancourt et al., 2022)	Random Forest	Land cover, agriculture, ozone, population, light pollution	Feature engineering, feature selection, spatial cross validation, Shapely Additive Explanations	RMSE, R2 score	Model performs average. Main limitation is that dataset is small and only one pollutant is used.
2	(Sanjeev , 2021)	RF, ANN, SVM	CO, PM_{10}, NO_2,NO_x, O_3,$PM_{2.5}$, SO_2, PM_{10}	Data preprocessing, normalization, attribute selection	Recall, Precision, F-score, specificity	RF outperformed other models with an accuracy of 99.4%
3	(Gopalakrishnan , 2021)	Ridge regression, LR, Elastic Net, RF, XGBoost	NO_2, Black carbon	Feature engineering, Correlation		Proposed model forecasts the BC and NO_2 concentration in Oakland area
4	(Harishkumar et al., 2020)	LR, RF, XGBoost, KNN, DT, ANN	$PM_{2.5}$	Cross validation	RMSE, MAE, MSE, R2	XGBoost outperformed all other models
5	(Zamani Joharestani et al., 2019)	RF, XGBoost, Deep Learning	$PM_{2.5}$, satellite and meteorological data, geographical data,	Aerosol Optical Depth	R2, MAE, RMSE	XGBoost performed best. Inclusion of satellite derived aerosol optical depth did not improve the accuracy
6	(Bhalgat et al., 2019)	ANN, Kriging	SO_2, $PM_{2.5}$	Data preprocessing, AR, ARIMA	MSE	SO_2 concentration is deadly in Nagpur and it is gradually increasing in Pune and Mumbai
7	(Zhu et al., 2018)	Baseline model, heavy model, light model	O3, SO_2, $PM_{2.5}$	Standard Frobenius norm regularization, nuclear norm	RMSE	Proposed light formulation outperforms other two model formulations. Regularization also

				regularization, l-2,1-norm regularization, consecutive close regularization		helps to boost the performance of the predictors.
8	(J. Zhang & Ding, 2017)	Feed forward Neural Network-back propagation, Extreme Learning Machine	NO_2,NO_x, O_3,$PM_{2.5,}$ SO_2	10-fold cross-validation	MAE, RMSE, R^2 , IA	ELM outperformed based on precision, robustness and generalization.
9	(Asgari et al., 2017)	Multinomial Logistic Regression and Multinomial Naïve Bayes algorithm	CO, PM_{10}, NO_2,NO_x, O_3,$PM_{2.5,}$ SO_2, temperature, pressure, cloud cover, relative humidity, wind speed, wind direction	Inverse distance weighting method	Precision, recall, F1 score	The overall accuracy of the ML algorithms are acceptable but not able to predict the minority classes properly due to the imbalanced dataset.
10	(Kleine Deters et al., 2017)	Proposed machine learning model	$PM_{2.5}$, meteorological data	Optimization	MSE, MAPE,	regression analysis provides better prediction of PM2.5 if the climatic conditions are extreme like high precipitation levels or strong winds

Gaussian distribution: Data is stationary if it follows Gaussian distribution. The histogram of the data is plotted in Fig 4 and it shows Gaussian distribution indicating the stationarity of data.

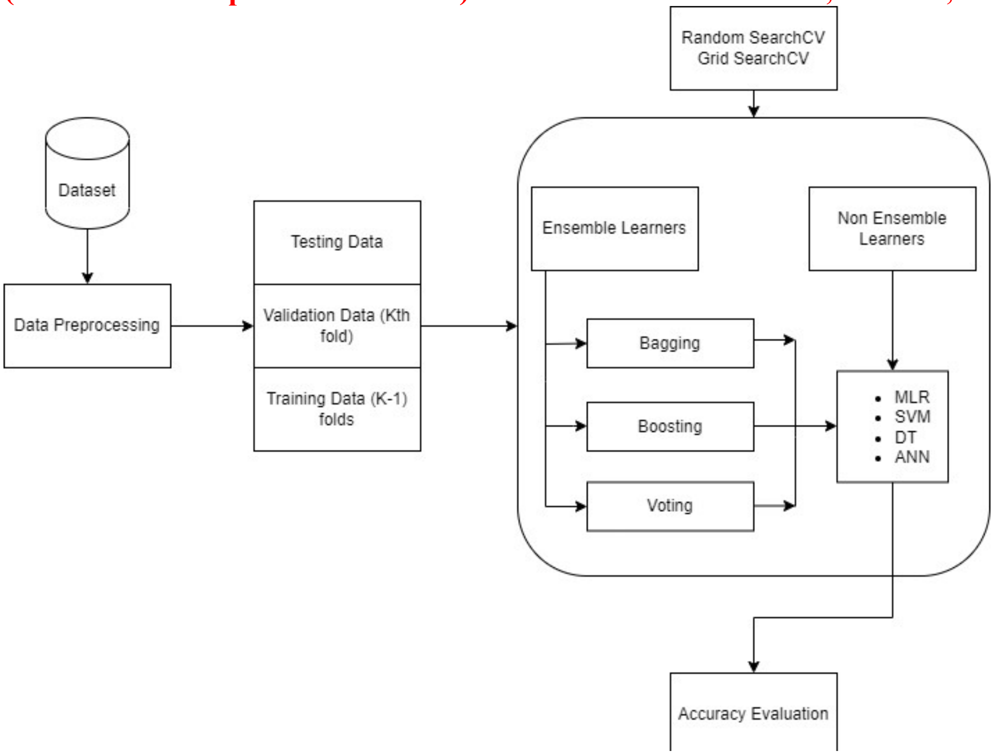


Fig 1. Proposed Study Framework

Table 3. Summary of measurement site and observed variables

Measurement Site	**Type**	**Variables**
Guwahati city	Meteorological conditions	Relative humidity
		Wind speed
		Wind direction
		Temperature
		Rainfall
	Criteria gases	NO_2
		SO_2
	Particulates	PM2.5
		PM10

Summary statistics: The data is partitioned into two intervals and checked for obvious and significant differences in summary statistics. The mean and variance of the two partitions are shown in Table 5. From the Table 5 it can be seen that the mean and variance of the two partitioned data is almost same indicating stationarity of data.

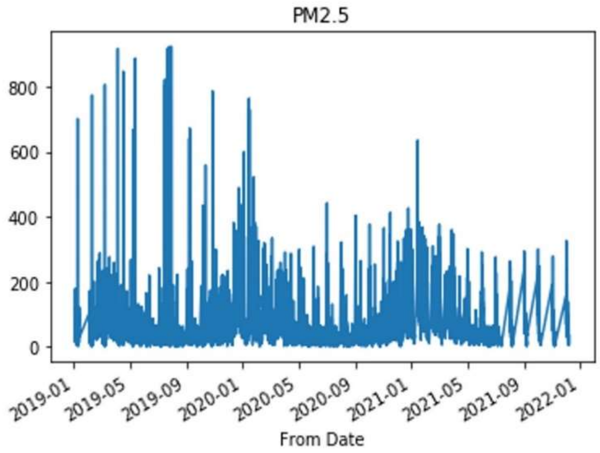
Data preprocessing: The performance of a machine learning algorithm is often influenced by the data preprocessing step (Wilson & Suh, 1997). Good quality data is the most important prerequisite for performing effective visualization and prediction using machine learning models. Data preprocessing helps to reduce the noise present in the dataset which helps to increase the processing speed and generalization capability of the machine learning algorithms. Two most important issues that we need to deal with are missing values and outliers. We have filled out *not-a-number* (NAN) data, removed outlier data. The cleaning process is applied to the data needs to be analyzed. Missing values of each of the feature in the dataset is shown in Fig 3. It can be observed that RF has highest number of missing values and RH, SR and BP has least number of missing values. A variety of factors may be responsible for missing values such as fault in the sensory device, manual error, etc. An imputer function is used to perform the process of interpolation. The strategy that is used here is mean value. No missing values are obtained after performing the interpolation process.Inter Quantile Range (IQR) is used to detect the outliers. Quantile based flooring and capping is used to deal with the outliers. The boxplot for outliers is shown in Fig 5.The data contains multiple inputs having different units. It is important that all the data are scaled into a particular range so that all attributes get equal weightage. Normalization is done so that an attribute having lesser significance with a large scale doesn't suppress another attribute of greater significance.

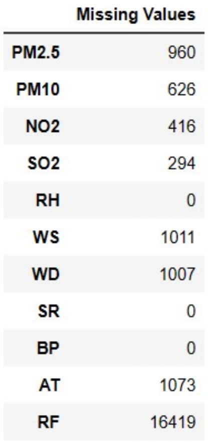
Table 4. Dataset descriptive statistics

	PM2.5	**PM10**	**NO2**	**SO2**	**WS**	**WD**	**AT**	**RF**
count	20214	20214	20214	20214	20214	20214	20214	20214
mean	63.12	121.26	14.14	18.01	1.12	151.80	24.50	0.07
std	68.41	134.78	12.39	5.83	0.73	58.32	4.87	0.22
min	0.06	0.50	0.02	3.91	0.03	13.95	9.29	0.00
25%	19.00	34.27	6.47	13.47	0.56	102.81	21.50	0.00
50%	41.11	75.25	10.60	17.14	0.99	139.10	24.82	0.00
75%	81.89	154.69	16.99	22.41	1.50	192.35	27.86	0.00
max	923.08	1000.00	107.04	140.84	23.71	323.83	37.64	3.67
skew	3.29	2.73	2.55	1.64	2.93	0.66	-0.39	6.14
kurt	21.19	10.53	8.89	15.34	49.01	-0.66	-0.25	56.41
var	4681.2	18168.2	153.5	33.9	0.5	3402.0	23.78	0.05

MinMaxScaler is used for normalization. It subtracts the minimum value from the attribute and then divides it by the range. The difference is defined by the difference between the maximum and the minimum value. The mathematical formula used to normalize the dataset is given in Eq.2(*Feature Scaling*, n.d.)

$$\text{Xscaled} = \frac{X-Xmin}{Xmax-Xmin} * (D-C) + C \qquad (2)$$





	Missing Values
PM2.5	960
PM10	626
NO2	416
SO2	294
RH	0
WS	1011
WD	1007
SR	0
BP	0
AT	1073
RF	16419

Fig 2. Time plots of the data**Fig 3**: Missing values of the features

Feature selection

Feature selection is the process of selecting a subset of attributes from the dataset that contains the most relevant information for performing the prediction. Researchers suggest that reducing the number of input variables helps to lower the computational cost of modelling and hence improves the prediction capacity(Medhi et al., 2016).Correlation matrix is used to check for correlation between features and hence determine the optimal number of input variables. It can be seen in Fig 6 that the attributes are not highly correlated. Correlation is very less.So, all the attributes present in the dataset is considered. Feature extraction is performed if there is redundant data.It involves selecting the optimum attributes. It is observed that many machine learning models performs better when they have a normal distribution.

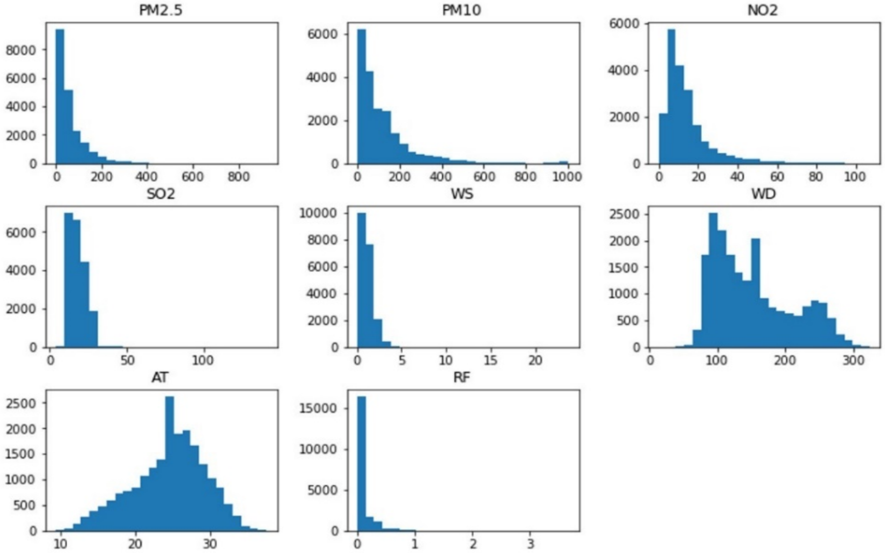
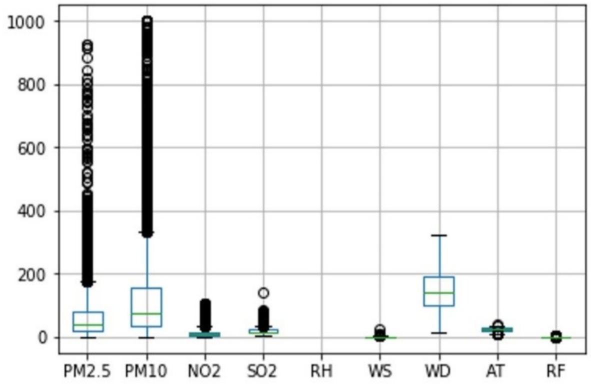


Fig 4. Histogram of data

Table 5. Mean and variance of partitioned data

	Mean	**Variance**	
Partition 1	49.36	6506.42	
Partition 2	49.15	5971.78	

They underperform when skewness is found. It is therefore, important to identify if skewness is present in the data and perform transformations and mappings to convert the skewed distribution to normal distribution. Fig 6 shows the skewness values of different features. It is observed that RF has the highest skewness and AT has the lowest skewness. Logarithmic transformation is used to reduce the skewness.



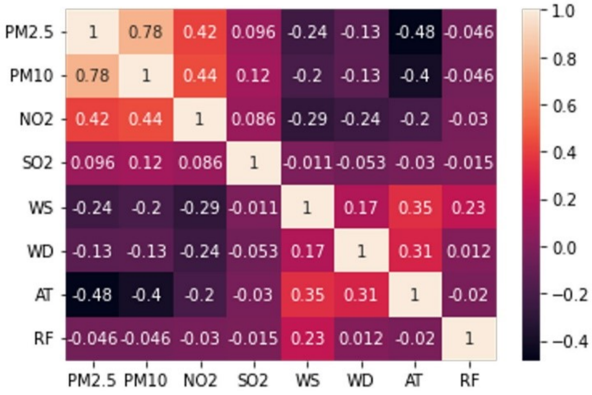


Fig 5. Detection of outliers**Fig 6**. Correlation matrix

Model training: Scikit-Learn library and Python library was used to build 16 models: (4) models using non-ensemble machine learning, (5) using bagging, (4) using boosting and (4) using voting to predict PM2.5 concencentration. Machine learning is the science by which machines can be made to act without programming them explicitly. Hyperparameter tuning is fitted into the models in order to reduce the error rate. We have used RandomSearchCV and GridSearchCV to perform hyperparameter tuning. In order to obtain an enhanced evaluation of training accuracy, 10-fold cross validation (10-CV) was done. Using 10-CV method the training dataset is divided into 10 subsets. Out of the 10 subsets, 9 are used for training each model and 1 subset is used as testing dataset. This process is repeated 10 times representing ten folds in 10-CV. The base learner parameters were as follows: ANN uses Multi -Layer Perceptron with three hidden layers: 5, 5 and 10 nodes respectively. Rectified Linear Unit (ReLU) is used as the activation function and adam is used as optimizer. Maximum iteration was set to 2000. For SVM, RBF kernel is used and regularization of 100 is used. For DT, criterion=MSE is used.To evaluate the model performance between ensemble and non-ensemble machine learning models, five evaluation metrics were used, namely, MAE, R^2, MSE, RMSE and RMSLE.

5. RESULTS AND DISCUSSION

In this section we have discussed the experimental design and empirical analysis for the prediction of PM2.5. Before applying the techniques, dataset is split into two parts: training set 70% and testing set 30%. From Fig 7, it can be seen that the average PM2.5 concentration is high during weekdays and

becomes low during weekends. This may be due to the use of more vehicles and running of factories due to workdays. It is also observed that during winters the PM2.5 concentration is highest and during summer it is least. The main reason for more air pollution during winter is due to winter inversion. During winter season, the atmospheric air of the earth becomes denser and cooler. Inversion process takes place where cool air is trapped by warm air, hence, forming a type of atmospheric

lid.

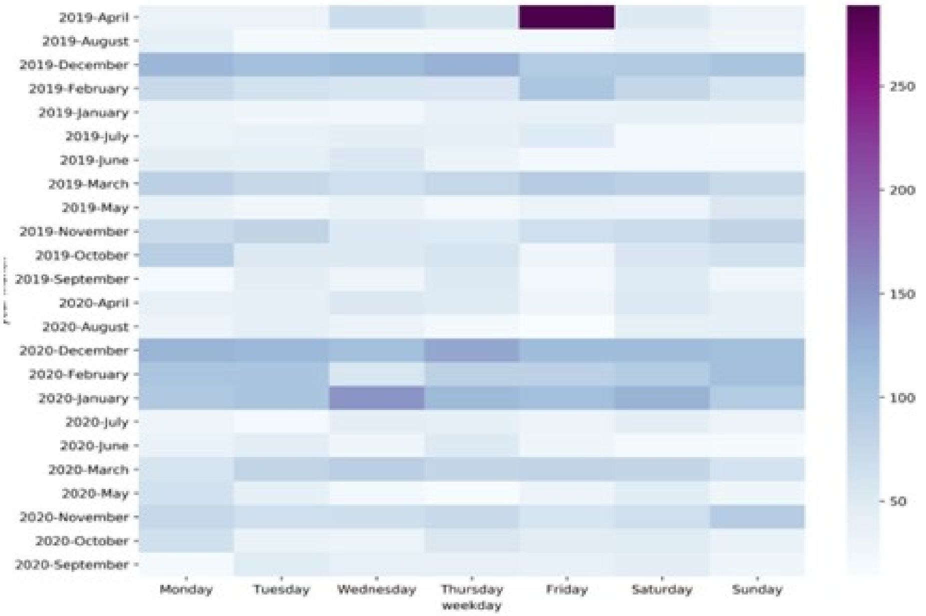


Fig. 7. Monthly and weekly PM2.5 concentration in Guwahati city

A range of experiments were carried out from 1 to 20 hidden layers to determine the optimal number of hidden layers that needs to be used for ANN. In Fig 8, the accuracy attained at different hidden layers is shown. R^2=0.99 is the highest value that is achieved. Three hidden layers were selected because error values obtained were considerably low (MAE=0.009, MSE=1.10-e04, RMSE=0.001) when compared with other hidden layers with same R^2 score. R^2 score for each of the model is calculated using hyperparameter tuning (HT) and without using hyperparameter tuning. For MLR hyperparameter tuning is not done. As shown in Fig 9 (a), the performance of ANN is best in both scenarios, namely, without hyperparameter tuning (0.97) and with hyperparameter tuning (0.99). Correlation coefficient of SVM is lowest among all the models with (0.72) without hyperparameter tuning and (0.75) with hyperparameter tuning. The performance of ANN is followed by MLR and then DT. If we compare both the scenarios, the overall performance of all the models have improved when hyperparameter tuning is done.

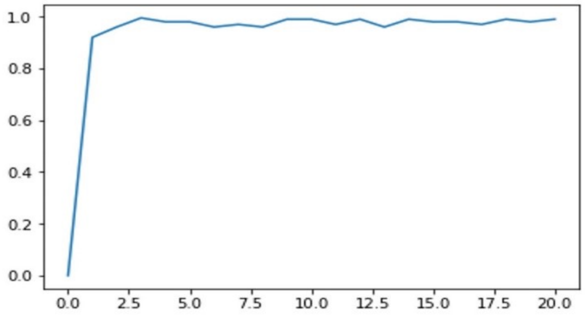
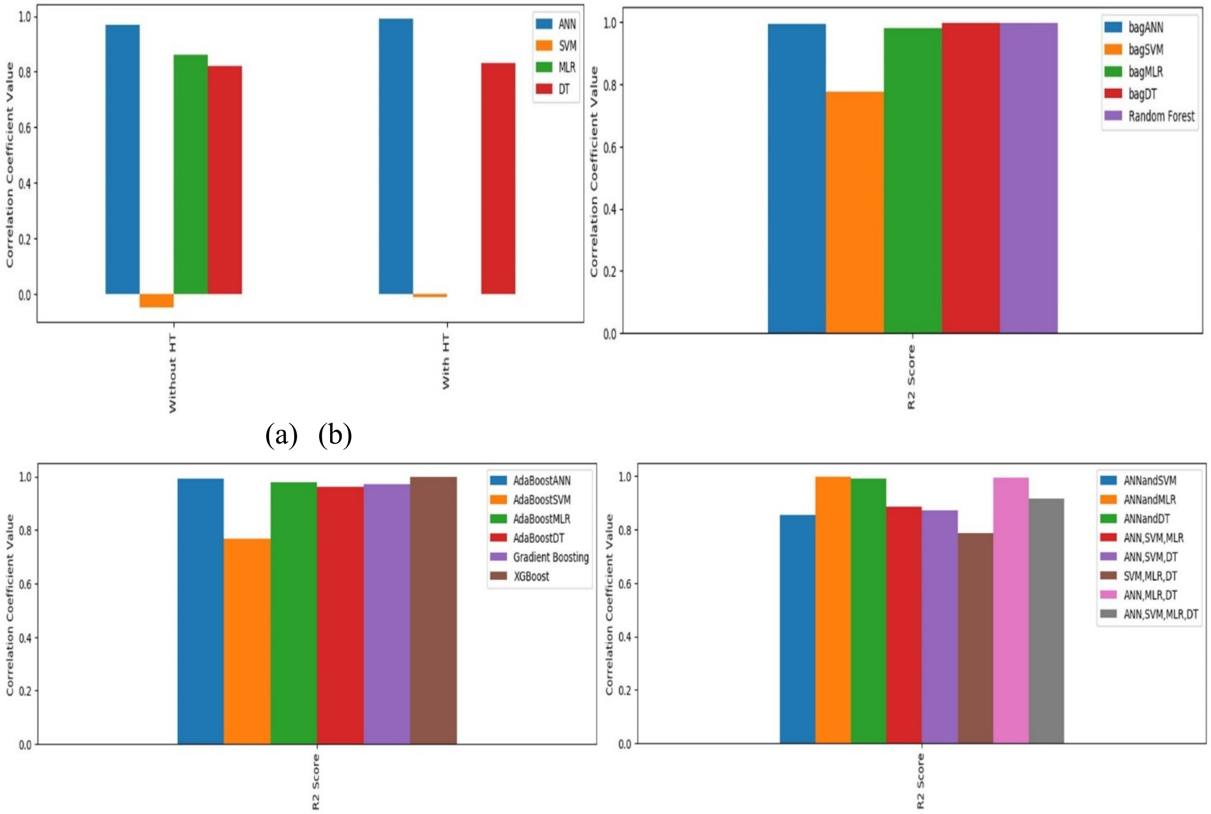


Fig 8. Correlation coefficient to select hidden layers for ANN



(a) (b)

(c) (d)

Fig 9. R^2 score of (a)non-ensemble machine learning models with and without hyperparameter tuning
(b)bagged ensemble machine learning models (c)boosting ensemble machine learning models
(d)voting ensemble machine learning models

From Fig 9(b), correlation performance reveals that the performance of bagging performed using ANN (bagANN), MLR (bagMLR) and DT (bagDT) is almost same with some minute difference in performance with hyperparameter tuning and without hyperparameter tuning. Performance of bagging by using SVM (bagSVM) still remains low. It is also interesting to notice that the correlation coefficient of non-ensemble ANN, MLR, SVM and DT improved upon the application of ensemble learning. Out of all the bagged models the best performance was given by bagDT (0.999). The real vs predicted PM2.5 value is shown in Fig 17. In the second part of the analysis, comparative analysis of all the non-ensemble machine learning models are done individually. The prediction performance of the models with hyperparameter and without hyperparameter is shown in Table 6. After comparing the evaluation metrics of all the non-ensemble machine learning algorithms in Table 6, it can be observed that SVM has the worst prediction performance. The prediction performance of ANN, MLR and DT are significantly good with ANN performing the best. It is shown in Fig 10(a).

Table 6. Prediction performance of non-ensemble machine learning algorithms

Non-ensemble Model	Without hyperparameter tuning					With hyperparameter tuning				
	R^2	MAE	MSE	RMSE	RMSLE	R^2	MAE	MSE	RMSE	RMSLE
ANN	0.9832	0.0231	1.8744e-05	0.0026	0.0084	0.9966	0.0033	1.9191e-05	0.0005	0.0044
SVM	-0.0595	0.7853	0.0089	0.1784	0.0972	-0.0195	0.0732	0.00634	0.0834	0.0561
MLR	0.8686	0.0027	5.4552e-06	0.0034	0.0025					
DT	0.829	0.0862	9.2318e-06	0.0089	0.0098	0.8391	0.0094	8.6734e-06	0.0056	0.0067

Different ensemble techniques, namely, bagging, boosting and voting were applied using the weak learners (ANN, SVM, MLR, DT). A comparative analysis is done on both the non-ensemble and ensemble machine learning techniques. The prediction performance of various ensemble techniques is shown in Table 7, 8 and 9. The prediction performance of bagging approach is shown in Table 7. It can be seen that the R^2 score has improved and MAE, RMSE, MSE values have decreased considerably than the models used individually. bagSVM has improved its prediction performance when compared to its corresponding value in Table 6 when used as a non-ensemble algorithm. Still bagSVM has performed weakly as compared to other bagged models. It is observed that Random Forest performed best with R^2 score of 0.9982 which is almost close to 1.00 and has the lowest error evaluation metrics of MAE=0.0043, MSE=6.874e-06 and RMSE=0.0047. It is shown in Fig 10(b). BagDT performs better than bagANN, bagMLR and bagSVM. It is shown in Fig 10(e).

In Table 8, a comparative analysis is done on the 6 models that are created using boosting ensemble approach. All the models performed better when hyperparameter tuning is applied. It is observed that XGBoost outperformed all the other models with R^2=0.9997, MAE=0.0025, MSE=3.4571e-05, RMSE=0.0025. It is also observed that boosting performed better than boosting when the same base learners were taken. It is shown in Fig 10(c). As we can see in Table 9, the predictive performance is further improved when voting approach is used. One important thing to note is that SVM which performed very weakly when used as non-ensemble algorithm or in bagging or boosting, its performance improved significantly when combined with other methods in voting. Voting when combined with ANN and MLR outperformed all other models. It is shown in Fig

10(d).A rigorous comparison is performed between non-ensemble (ANN, SVM, MLR, DT) and ensemble (bagging,
boosting, voting) machine learning techniques to predict PM2.5 concentration in the city of Guwahati.MAE, MSE, RMSE and RMSLE were used to measure the response variable in absolute terms. When compared to the non-ensemble models, the error evaluation metrics values were much lower in the ensemble machine learning models. Similarly, the R^2 score of ensemble methods are much higher than the non-ensemble machine learning models. This indicates that ensemble models are reliable and consistent if we compare with their respective single non-ensemble models. The results obtained during testing and training using bagging ensemble approach reveals that it has better capabilities in displaying the uncertainties embedded in the data as it can be seen that the values of correlation coefficient significantly improved if we compare with other learners.

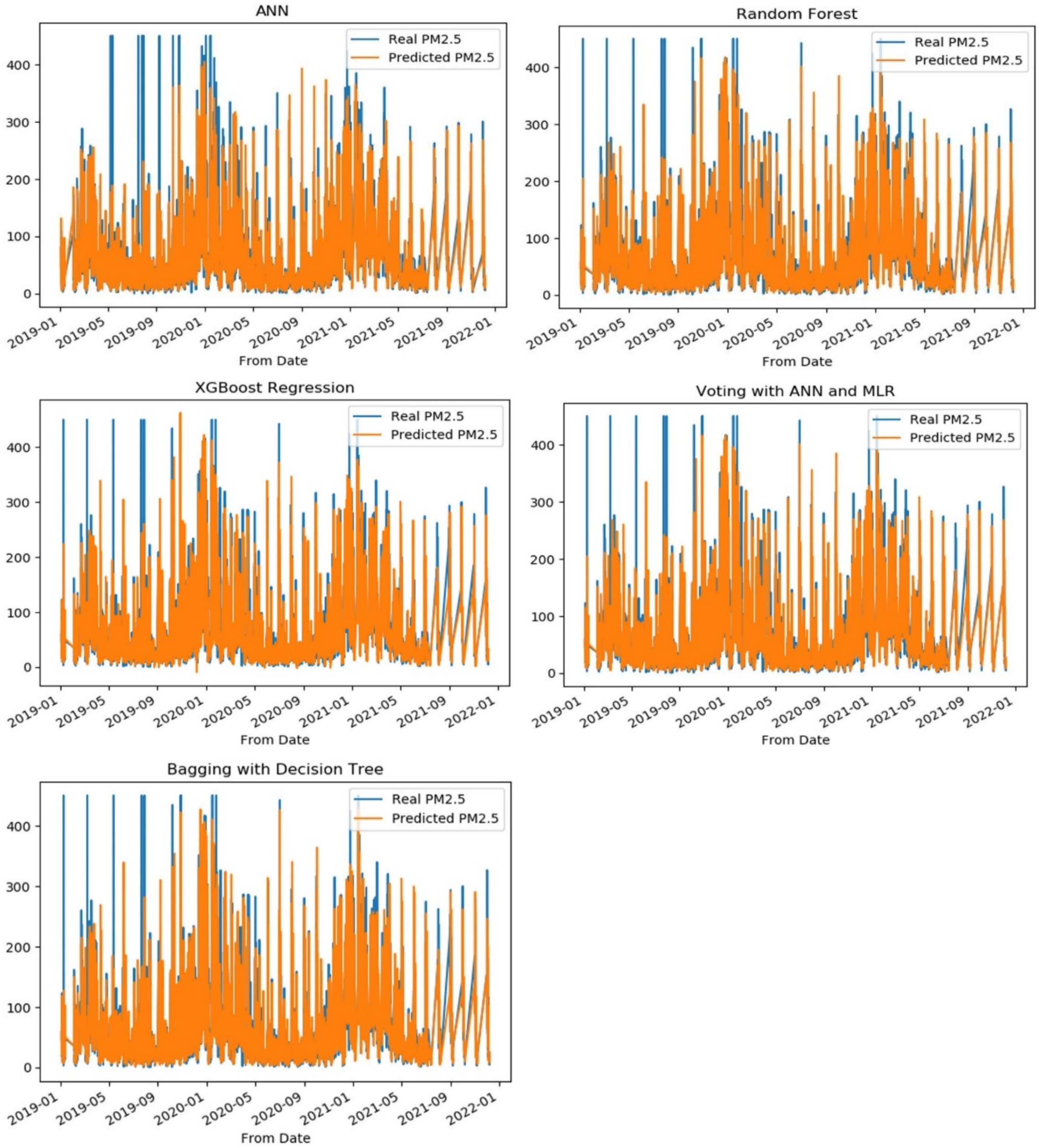


Fig 10. Real Vs Predicted PM2.5 valueusing (a)ANN (b)XGBoost (c)Random Forest (d)voting with ANN and MLR
(e)bagging with DT

6. CONCLUSION

Further, SVM when used as a single model or with bagging or boosting performed very weakly. But through the use of voting ensemble technique the performance of SVM increased significantly. This proves that by applying ensemble methods we can improve the performance of weak learners into strong ones by the combination of the appropriate ensemble algorithms. In general we can conclude that ensemble techniques have great prospect in improving the prediction accuracy of single base learners. The limitation of the work is that few extra features like traffic and emission from industries should also be considered.

Table 7. Prediction performance of bagging approach

Ensemble Model	R^2	MAE	MSE	RMSE	RMSLE
bagANN	0.9935	0.0056	3.4317e-05	0.0005	0.0048
bagSVM	0.7765	0.0521	0.0021	0.0521	0.0545
bagMLR	0.9825	0.0081	9.2567e-06	0.0134	0.0095
bagDT	0.9962	0.0054	4.0671e-05	0.0063	0.0052
Random Forest	0.9982	0.0043	6.874e-06	0.0047	0.0038

Table 8. Prediction performance of boosting approach

Ensemble Model	R^2	MAE	MSE	RMSE	RMSLE
AdaBoostANN	0.9992	0.0026	3.1651e-06	0.0024	0.0021
AdaBoostSVM	0.7684	0.0856	0.0067	0.0831	0.0634
AdaBoostMLR	0.9786	0.0074	0.0002	0.0218	0.0096
AdaBoostDT	0.9618	0.0082	0.0006	0.0245	0.0092
Gradient Boosting	0.9742	0.0075	0.0002	0.0234	0.0093
XGBoost	0.9997	0.0025	3.4571e-05	0.0025	0.0026

Table 9. Prediction performance of voting approach

Ensemble Model	R^2	MAE	MSE	RMSE	RMSLE
ANN, SVM	0.8561	0.0425	0.0024	0.0467	0.0356
ANN, MLR	0.9984	0.0025	5.1263e-06	0.0035	0.0025
ANN, DT	0.9925	0.0046	3.2377e-05	0.0062	0.0053
ANN, SVM, MLR	0.8856	0.0361	0.0002	0.3281	0.0262
ANN, SVM, DT	0.8727	0.0354	0.0008	0.0363	0.0365
SVM, MLR, DT	0.7856	0.3681	0.0006	0.0367	0.0351
ANN, MLR, DT	0.9949	0.0027	2.5684e-05	0.0049	0.0042
ANN, SVM, MLR, DT	0.9162	0.0256	0.0005	0.0272	0.0245

BIBLIOGRAPHY:

Alade, I. O., Abd Rahman, M. A., & Saleh, T. A. (2019). Predicting the specific heat capacity ofalumina/ethylene glycol nanofluids using support vector regression model optimized with Bayesian algorithm. *Solar Energy*, *183*, 74–82.

Asgari, M., Farnaghi, M., &Ghaemi, Z. (2017). Predictive mapping of urban air pollution using Apache Spark on a Hadoop cluster. *Proceedings of the 2017 International Conference on Cloud and Big Data Computing*, 89–93.

Barman, N., & Gokhale, S. (2019). Urban black carbon-source apportionment, emissions and long-range transport over the Brahmaputra River Valley. *Science of the Total Environment*, *693*, 133577.

Betancourt, C., Stomberg, T. T., Edrich, A.-K., Patnala, A., Schultz, M. G., Roscher, R., Kowalski, J., &Stadtler, S. (2022). Global, high-resolution mapping of tropospheric ozone–explainable machine learning and impact of uncertainties. *Geoscientific Model Development Discussions*, 1–36.

Bhalgat, P., Bhoite, S., &Pitare, S. (2019). Air quality prediction using machine learning algorithms. *International Journal of Computer Applications Technology and Research*, *8*(9), 367–370.

Bougoudis, I., Demertzis, K., &Iliadis, L. (2016). HISYCOL a hybrid computational intelligence system for combined machine learning: The case of air pollution modeling in Athens. *Neural Computing and Applications*, *27*(5), 1191–1206.

Evans, J., van Donkelaar, A., Martin, R. V., Burnett, R., Rainham, D. G., Birkett, N. J., &Krewski, D. (2013). Estimates of global mortality attributable to particulate air pollution using satellite imagery. *Environmental Research*, *120*, 33–42.

Feature Scaling: Normalization and Standardization - Quinn-Yann - 博客园. (n.d.). Retrieved January 11, 2023, from https://www.cnblogs.com/quinn-yann/p/9808247.html

Gonzalez-Gorman, S., Kwon, S.-W., & Patterson, D. (2019). Municipal efforts to reduce greenhouse gas emissions: Evidence from US cities on the US-Mexico border. *Sustainability*, *11*(17), 4763.

Gopalakrishnan, V. (2021). *Hyperlocal air quality prediction using machine learning. Towards data science*.

Harishkumar, K. S., Yogesh, K. M., & Gad, I. (2020). Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models. *Procedia Computer Science*, *171*, 2057–2066.

Hsieh, H.-P., Lin, S.-D., & Zheng, Y. (2015). Inferring air quality for station location recommendation based on urban big data. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 437–446.

Janssen, N. A. H., Fischer, P., Marra, M., Ameling, C., &Cassee, F. R. (2013). Short-term effects of PM2. 5, PM10 and PM2. 5–10 on daily mortality in the Netherlands. *Science of the Total Environment*, *463*, 20–26.

Johnson, M., Isakov, V., Touma, J. S., Mukerjee, S., &Özkaynak, H. (2010). Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmospheric Environment*, *44*(30), 3660–3668.

Kim, K.-H., Kabir, E., & Kabir, S. (2015). A review on the human health impact of airborne particulate matter. *Environment International*, *74*, 136–143.

Kioumourtzoglou, M.-A., Schwartz, J. D., Weisskopf, M. G., Melly, S. J., Wang, Y., Dominici, F., &Zanobetti, A. (2016). Long-term PM2. 5 exposure and neurological hospital admissions in the northeastern United States. *Environmental Health Perspectives*, *124*(1), 23–29.

Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., &Rybarczyk, Y. (2017). Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters. *Journal of Electrical and Computer Engineering*, *2017*.

Laden, F., Schwartz, J., Speizer, F. E., & Dockery, D. W. (2006). Reduction in fine particulate air pollution and mortality: Extended follow-up of the Harvard Six Cities study. *American Journal of Respiratory and Critical Care Medicine*, *173*(6), 667–672.

Li, X., Peng, L., Hu, Y., Shao, J., & Chi, T. (2016). Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, *23*(22), 22408–22417.

Medhi, S., Ahmed, C., & Gayan, R. (2016). A study on feature extraction techniques in image processing. *International Journal of Computer Sciences and Engineering*, *4*(7), 89–93.

Medhi, S., &Gogoi, M. (2021). Visualization and Analysis of COVID-19 Impact on PM2. 5 Concentration in Guwahati city. *2021 International Conference on Computational Performance Evaluation (ComPE)*, 012–016.

Pohjola, M. A., Kousa, A., Kukkonen, J., Härkönen, J., Karppinen, A., Aarnio, P., &Koskentalo, T. (2002). The spatial and temporal variation of measured urban PM10 and PM2. 5 in the Helsinki metropolitan area. *Water, Air and Soil Pollution: Focus*, *2*(5), 189–201.

Pope III, C. A., & Dockery, D. W. (2006). Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air & Waste Management Association*, *56*(6), 709–742.

Sanjeev, D. (2021). Implementation of machine learning algorithms for analysis and prediction of air quality. *Int J Eng Res Technol*, 2278–0181.

Sharma, S. (2021, September 27). What is Air Quality Index (AQI) & How Is It Calculated ?*Prana Air*. https://www.pranaair.com/blog/what-is-air-quality-index-aqi-and-its-calculation/

Wang, J., & Ogawa, S. (2015). Effects of meteorological conditions on PM2. 5 concentrations in Nagasaki, Japan. *International Journal of Environmental Research and Public Health*, *12*(8), 9089–9101.

Wilson, W. E., & Suh, H. H. (1997). Fine particles and coarse particles: Concentration relationships relevant to epidemiologic studies. *Journal of the Air & Waste Management Association*, *47*(12), 1238–1249.

Xi, X., Wei, Z., Xiaoguang, R., Yijie, W., Xinxin, B., Wenjun, Y., &Jin, D. (2015). A comprehensive evaluation of air pollution prediction improvement by a machine learning method. *2015 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, 176–181.

Yan, C., Xu, S., Huang, Y., Huang, Y., & Zhang, Z. (2017). Two-phase neural network model for pollution concentrations forecasting. *2017 Fifth International Conference on Advanced Cloud and Big Data (CBD)*, 385–390.

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., &Talebiesfandarani, S. (2019). PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere*, *10*(7), 373.

Zhang, F., Cheng, H., Wang, Z., Lv, X., Zhu, Z., Zhang, G., & Wang, X. (2014). Fine particles (PM2. 5) at a CAWNET background site in Central China: Chemical compositions, seasonal variations and regional pollution events. *Atmospheric Environment*, *86*, 193–202.

Zhang, J., & Ding, W. (2017). Prediction of air pollutants concentration based on an extreme learning machine: The case of Hong Kong. *International Journal of Environmental Research and Public Health*, *14*(2), 114.

Zhu, D., Cai, C., Yang, T., & Zhou, X. (2018). A machine learning approach for air quality prediction: Model regularization and optimization. *Big Data and Cognitive Computing*, *2*(1), 5.